XSEDE Data Management Use Cases

March 20, 2014

Version 1.6

Table of Contents

A Document History

Overall Document Authors:

Christopher Jordan
Texas Advanced Computing Center
The University of Texas at Austin
Austin, TX
ctjordan@tacc.utexas.edu

Kurt Seiffert
Indiana University
Bloomington, Indiana
seiffert@iu.edu

Altaf Hossain
Pittsburgh Supercomputing Center (PSC)
Pittsburgh, PA
altaf@psc.edu

Ian Foster
The University of Chicago and Argonne National Laboratory
Argonne, IL 60439
foster@anl.gov

Morris Riedel Jülich Supercomputing Centre
Forschungszentrum Jülich GmbH
D-52425 Jülich
Germany

Andrew Grimshaw
University of Virginia
PO Box 400740
Charlottesville VA 22904
grimshaw@virginia.edu

David Lifka
Cornell University
512 Frank H. T. Rhodes Hall
Ithaca, NY 14853
lifka@cac.cornell.edu

Description	Version	Date	Changes	Author
First formatted and accepted version.	1.5	4/23/2013	Full text.	Altaf Hossain, Chris Jordan
Revised and added UCDM 5.0	1.6	10/30/2013	Added use case #5.	Altaf Hossain, Chris Jordan

B Document Scope

This document is both a user-facing document (publically accessible) and an internal working document intended to define user needs and use cases that fall under the general umbrella of Data Management within the overall activities of XSEDE. The definition of use cases is based on a template from Malan and Bredemeyer[1]. In general it is in keeping with the approaches and philosophy outlined in "Software architecture in practice."[2]

This document is one component of a process that generates at least the following documents, some of which are user-facing, some are as of now intended to be internal working documents:

- ***This document*** - A description of use cases [User facing]
- A binary mapping of use cases to Requirements in DOORS (a binary mapping – for each use case a "yes" or "no" flag indicating whether a particular requirement within the full list of requirements is or is not required to enable a particular use case
- A set of level 3 decomposition documents, which include:
    - Quality Attributes descriptions
    - Connections diagram in UML

The use cases are presented here using the following format, derived from the Malan and Bredemeyer white paper[1] as follows:

Use Case	Use case identifier and reference number and modification history
Description	Goal to be achieved by use case and sources for requirement
References	References and citations relevant to use case
Actors	List of actors involved in use case

[1] Malan, R., and D. Bredemeyer. 2001. Functional requirements and use cases. *www.bredemeyer.com/pdf_files/functreq.pdf*
[2] Bass, L., P Paul Clements, and Rick Kazman

Prerequisites (Dependencies) & Assumptions	Conditions that must be true for use case to be possible Conditions that must be true for use case to terminate successfully
Steps	Interactions between actors and system that are necessary to achieve goal
Variations (optional)	Any variations in the steps of a use case
Quality Attributes	
Non-functional (optional)	List of non-functional requirements that the use case must meet
Issues	List of issues that remain to be resolved

C Summary of Data Management Use Cases

The Data Management use cases cover a broad range of activities related to data movement, storage, and archival by users and communities of users in the context of XSEDE. These include interactions with data collections created outside of XSEDE, data collections created using XSEDE resources, and communities and collections crossing the boundary of XSEDE to include campus-level resources. The current use cases cover basic data sharing by a user community (UCDM1.0), execution of a coordinated program of research using a shared data collection (UCDM2.0), and a coordinated program of research utilizing data generated by a sensor or network of sensors on a regular or continuous basis (UCDM3.0). Future use cases will address migration of a users data in association with a change of primary computational activity or a change of service provider status (UCDM 4.0), and creation, access, and search of metadata associated with a data collection (UCDM5.0).

D Data Management Use Cases

Use Case	**UCDM1.0: Share and manage a common repository of data with a distributed user community**
Description	A group of users, individually or collectively, build a collection of data and software code, which must be shared with and managed by a community of

	users using many XSEDE resources.
References	References and citations relevant to use case: TBD.
Actors	Collection Manager: A user with elevated privileges to deposit and manage data in the collection. Collection User: A user with limited privileges to access data and code in the collection. User Agent: A script or function that acts on behalf of a user on one or more XSEDE resources. Collective-User Agent: A script or function that acts on behalf of the group of users, typically to perform regular curation tasks. Data gateway: A system or service that provides open access to a shared collection of data to users outside of the set of XSEDE resources
Prerequisites (Dependencies) & Assumptions	All users of the shared data collection have XSEDE or other identities that can be used to set and enforce permissions. Users will have differing levels of privilege to create and/or alter data within the collection. All computational and login systems on which users access data are within the XSEDE "circle of trust" and use the XSEDE accounting system to exchange information regarding usage, accounts, and allocations.
Steps	1. Collection Manager logs in to an XSEDE resource 2. The Collection Manager moves (or copies) data from a local data resource into a shared resource using local data movement tools 3. The Collection Manager alters the structure of the collection (directory hierarchies, file locations) as appropriate for collective data sharing 4. The Collection Manager sets permissions on components of the collection to allow for public, semi-public, and private use of data by relevant actors 5. A Collection User logs in to (the same or a different) XSEDE resource 6. The Collection User copies data from the shared collection area into a local storage resource for use in computation or analysis tasks 7. The Collection user performs computation or analysis resulting in data for further sharing 8. The Collection user moves (or copies) their result data into the shared collection area for use by others
Variations (optional)	1. Instead of Collection Managers and users directly logging in to XSEDE resources to access data, they could use remote data transfer and/or access tools to retrieve and store data from the shared collection to campus-level resources for analysis and computational activities outside of XSEDE. 2. Some or all data may be available to any user without permissions restrictions.

Quality Attributes	When a failure or temporary interruption of network access to one or more XSEDE sites occurs, the shared data resource remains readable, to users at unaffected sites for up to 8-12 hours. Storage and retrieval of data can is not limited to a single user or site at any one time, i.e. parallel data access can occur both on a node and a site level. No collection- or project-specific software or tools are required to store or retrieve data from the shared collection.
Non-functional (optional)	
Issues	List of issues that remain to be resolved

Use Case	**UCDM2.0: Coordinated computation and analysis involving a shared body of data**
Description	A user or group of users executes a coordinated program of research involving multiple computational, analysis, and visualization tasks executing on multiple XSEDE resources and SPs, utilizing a common body of data.
References	References and citations relevant to use case: TBD.
Actors	Investigator: An individual user who initiates tasks as part of a coordinated research program User Agent: A script or function that acts on behalf of a user on one or more XSEDE resources.
Prerequisites (Dependencies) & Assumptions	All users/investigators have XSEDE or other identities by which permissions can be set and enforced. All users/investigators are part of a single XSEDE allocation that represents the program of research. All resources on which users access data are within the XSEDE "circle of trust" and use the XSEDE accounting system to exchange information regarding usage, accounts, and allocations.
Steps	1. A user logs in to an XSEDE resource 2. The user constructs a job making use of some shared data element accessible on all XSEDE resources 3. The user submits the job for execution on a specific XSEDE resource 4. The job executes, accessing the shared data element and producing output written to the common data location 5. The same or a different user logs in to a different resource 6. The user constructs a job making use of data produced as the output of step 4 7. The user submits the job for execution on a specific XSEDE resource 8. The job executes, reading in the output of step 4 and further

	producing output for users in subsequent tasks 9. Steps 5 through 8 are repeated by the same or different users
Variations (optional)	1. Instead of a specific resource, users may use the XSEDE user portal, a science gateway, or another indirect mechanism for constructing and submitting jobs for execution. 2. Some output data elements may be "open", i.e. not subject to permissions restrictions
Quality Attributes	When a failure or temporary interruption of network access to one or more XSEDE sites occurs, the shared data resource remains available to users at unaffected sites for 8-12 hours. Storage and retrieval of data can is not limited to a single user or site at any one time, i.e. parallel data access can occur both on a node and a site level.
Non-functional (optional)	
Issues	List of issues that remain to be resolved

Use Case	**UCDM3.0: Shared use of large-scale/streaming sensor input data**
Description	A group of users, individually or collectively, perform a set of computations or analyses using input data uploaded on a periodic or continuous basis from a sensor or system of sensors.
References	References and citations relevant to use case: TBD.
Actors	Sensor: A system making observations of some phenomena of research interest, such as a telescope, a weather station, or a seismograph Input gateway: A system enabling the storage of data from a sensor into an XSEDE data resource on a continuous basis User agent: A software agent acting on behalf of the user to retrieve and/or transform data deposited by the sensor on the user's behalf Input agent: A software agent automating the storage of data from a sensor into an XSEDE data resource on a continuous or periodic basis
Prerequisites (Dependencies) & Assumptions	All users of the sensor data have XSEDE accounts and are members of a single project or group used to define shared permission to access input data. The users or project managers have independently established access to one or more XSEDE resources for the purpose of automating data retrieval by the input gateway All resources on which users access data are within the XSEDE "circle of

	trust" and use the XSEDE accounting system to exchange information regarding usage, accounts, and allocations.
Steps	1. A user logs in to an XSEDE resource. 2. A user constructs a job making use of sensor input data, using variables or another mechanism to allow for selection of an input "window" in time or space. 3. The user submits and successfully tests the job execution with exemplary input data 4. The user makes this job available to others on the same or other resources. 5. An input agent retrieves new data from or through the input gateway, either continuously or on a regular, e.g. daily, basis. 6. One or more users reuse the job script from step 3 with the data from step 5. 7. The data resulting from step 6 may be further shared as in UCDM 1.0
Variations (optional)	1. Users may not always reuse the same job scripts with varying input data. Instead both the job and the input data may vary. 2. Jobs could be created and launched on behalf of a science gateway, or could be launched automatically to process data as it is created. 3. Jobs could be run on campus level resources with XSEDE services being used only to provide access to the shared data resource.
Quality Attributes	Successful storage of 95% or more of all sensor outputs successfully sent to the input agent (i.e. not experiencing a failure outside of XSEDE). Successful execution of 95% or more of tasks accessing the sensor data once it has been successfully retrieved and stored. Overall performance sufficient to retrieve results with a continuous generation rate of up to 10MB/sec.
Non-functional (optional)	
Issues	List of issues that remain to be resolved

Use Case	**UCDM4.0: Migration of data between resources**
Description	A user or project migrates all their active data to a new resource, either because of a change in allocation status or the introduction of a new resource into the XSEDE ecosystem.

References	References and citations relevant to use case: TBD.
Actors	Data owner: The "owner" or responsible party for a working set of data. User agent: A software agent acting on behalf of the user to copy or move data between resources and/or sites. Data Transfer service: a service or collection of services, running as a daemon on the source and destination resources, which provide data transfer functionality in conjunction with the user agent.
Prerequisites (Dependencies) & Assumptions	The user has existing allocations for resource access, and any required storage allocations, on both the source and destination resource. The data to be migrated is within reasonable parameters for the size of data relative to network capacity and target file system capacity.
Steps	1. The data owner identifies folders of importance to their ongoing efforts. 2. The data owner initiates a command or script to perform data transfer operations on a collection of files and/or folders. 3. Data is transferred by the user agent to the destination resource, and verification is performed. 4. When data transfer is complete for all identified data, notification is sent to the user of the completion of the transfer process.
Variations (optional)	1. There may be a period during which data is in use on two separate resources, and must be synchronized on a regular basis. 2. Identification of the data to be migrated could be automated based on storage allocation information, default migration of home directories, or another automatic selection process for data migration.
Quality Attributes	The data owner must perform no more than one initiation step per file system (or equivalent storage resource) from/to which data will be migrated. 100% of data must be successfully transferred between resources and verified. Data that fails to be transferred or to be verified post-transfer is retransferred automatically.
Non-functional (optional)	
Issues	List of issues that remain to be resolved

Use Case	UCDM5.0: Metadata Storage and Access
Description	Users create, manage, and search metadata attributes associated with a collection of data objects.
References	References and citations relevant to use case: TBD.
Actors	Data owner: The "owner" or responsible party for a working set of data. User agent: A software agent acting on behalf of a user to access and search metadata created by the data owner. Data Management Service: A service or services that provide management functions for data and metadata, which may include data transfer, but must include metadata storage and search, along with usable references to a service that does provide data transfer. Data Management Client: A software tool or service that provides mechanisms for registering, retrieving, and searching metadata on behalf of a user.
Prerequisites (Dependencies) & Assumptions	The data owner has existing allocations for the storage resource(s) on which the data and/or metadata are stored. If the data is not public, the user has an XSEDE account and permission to access metadata. Metadata formatting may vary significantly. For the purposes of this use case no specific format is assumed, and one or more free text or XML fields may be considered to satisfy the use case.
Steps	1. The data owner uploads one or more data objects into a storage resource, via the data management service. 2. The data management service creates a record of the object to which metadata can be attached. 3. The data owner manually, or via a software tool, creates or extracts metadata associated with the data object. 4. The data management client is invoked to store the metadata elements and associated them with the data object. 5. A user other than the data owner executes a data management client and connects to the data management service. 6. The user searches for specific metadata elements, or views all metadata associated with a specific object. 7. The user invokes a data transfer client, if different from the data management client, to retrieve a data object once its attributes have been confirmed via the metadata.
Variations (optional)	1. There may be multiple data "owners" who have permission to create or modify metadata associated with a data object

	2. Search mechanisms may vary significantly, but can be considered to include XML-based queries, SQL-based queries, or less structured "pure text" queries. 3. A web-based or other "front-end" interface may perform query and retrieval functions on behalf of a user or user group.
Quality Attributes	Metadata storage and retrieval commands must clearly indicate whether no metadata exists or no metadata was found that matched the parameters given. Metadata storage and search functions must be able to operate on large numbers of records within a reasonable time. (100 million per minute?) If multiple metadata servers/repositories are used to provide failover capability, there must be no more than 30 minutes differential in stored metadata.
Non-functional (optional)	
Issues	List of issues that remain to be resolved